Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2016) 24: 946 – 960

doi:10.3906/elk-1309-96

Quantitative information extraction from gas sensor data using principal component regression

Ahmet ÖZMEN[1], Bekir MUMYAKMAZ[2], Mehmet Ali EBEOĞLU[2,3], Cihat TAŞALTIN[3], İlke GÜROL[3], Zafer Ziya ÖZTÜRK[3,4], Deniz DURAL[1,*]

[1]Department of Computer Engineering, Sakarya University, Sakarya, Turkey
[2]Department of Electrical and Electronics Engineering, Dumlupınar University, Kütahya, Turkey
[3]TÜBİTAK, Marmara Research Center, Materials and Chemical Technologies Research Institute, Gebze, Kocaeli, Turkey
[4]Department of Physics, Gebze Institute of Technology, Gebze, Kocaeli, Turkey

Received: 12.09.2013 • **Accepted/Published Online:** 22.01.2014 • **Final Version:** 23.03.2016

Abstract: This paper presents a novel use of the principal component analysis (PCA) and regression methods for quantitative feature extraction from gas sensor data. In this approach, PCA plots are interpreted by observing the locations of samples in the principal component domain. A trainable data processing system that also produces numerical output is designed to validate the method. The main advantages of this system are: 1) retrainability: once it is trained, it can be used for any gas set; 2) flexibility: adaptation to different targets does not require hardware modifications (if a sufficient number and variety of sensors are installed in the sensor cell); and 3) simplicity: all computations are performed with only linear operators, and hence the system does not require complex structures or powerful computation resources.

Several experiments are conducted using two industrial gases (toluene and ethanol) to validate the approach. The new approach is also compared with two classic principal component regression (PCR) methods. The results show that the new approach performs better than the classic PCR approaches.

Key words: Electronic nose, principal component regression, quartz crystal microbalance, gas sensor data analysis

1. Introduction

Oscillation frequencies of quartz crystal microbalance (QCM) sensors change based on the mass of vapor absorbed by the film covering the quartz surface. QCM sensors are preferred in electronic nose (e-nose) applications for their linear response to concentrations over a wide range, from 100 ppm to 15,000 ppm. However, the complex circuitry for data acquisition and decreased sensitivity to lower concentrations (lower than 50 ppm) are the main drawbacks of QCM sensor systems.

Feature extraction from sensor data is an important task of e-nose systems. Depending on the application, information on qualitative and/or quantitative properties is extracted, determining the species or the concentration of a given species, respectively [1,2]. Since each sensor produces one-dimensional data (a concentration versus frequency shift), an array of sensors coated with different chemicals is used to detect various types. Species and concentration determination for any sample in the sample space covering all vapors by means of frequency shifts is quite difficult because, unlike the case with colors in digital pictures, there are no base

*Correspondence: ddural@sakarya.edu.tr

vectors in this case. Hence, linear and nonlinear feature extraction approaches are used to analyze QCM sensor data [3,4].

Principal component analysis (PCA) is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie in the first coordinate (called the first principal component), the second greatest variance lies in the second coordinate, and so on [5]. The PCA method is mostly used for classification of gas species by means of clustering. Sensor array data are transferred using PCA, and the clusters are identified on the PCA plot with their deviating locations. PCA plots visually help analyzers identify different aspects of smells (e.g., discriminating tea odor or coffee brands) [6,7]. The main objective of this type of work is to determine the cluster in which a sample resides. When principal component scores are used for estimating regression coefficients, the analysis method is called principal component regression (PCR) [8].

The general PCA approach is to extract qualitative information, such as the detection of gas species in terms of presence or absence. Obtaining quantitative information, however, is much more difficult and requires linear or nonlinear regression methods. In this work, a novel approach is developed for prediction of gas concentrations using PCA scores. Initially, the system is trained with samples collected from a target gas set. Once the system is trained, it can be used to estimate concentrations, providing both numerical and visual outputs. The numerical outputs are the ppm amounts of the constituent elements of gas mixtures. The visual output also helps a user infer the concentration from the location of the test sample in the PCA plot.

Simplicity and low computational complexity make this regression process valuable for mobile gas measurement systems due to limited resources. The proposed method in this study utilizes this property by using minimum necessary principal component variables and linear operators. Otherwise, classical regression methods that use whole sensors' data require much more memory and computational power. In the literature, there are other PCR methods that use nonlinear approaches such as artificial neural networks, but these methods are not suitable for mobile gas measurement systems [9,10].

2. QCM sensor systems

Fundamental to the e-nose system is the idea that each sensor in the array has a different sensitivity. For example, gas A may produce a high response in one sensor and a lower response in others, whereas gas B might produce high readings for sensors other than the one that is sensitive to gas A. Sensors are implemented with equivalent quartz crystals; however, different sensitivities are obtained by coating them with various chemicals.

The coating material is chemically formulated to shift the resonant frequency of a quartz sensor based on the mass change of a target gas. To produce the sensors, AT-cut quartz crystals with a 10 MHz fundamental frequency (Klove B.V., the Netherlands) are used. The jet-spray coating system is built up in a glove box to control the ambient temperature and humidity. The jet-spray pressure is controlled by a regulation valve. Solutions of the sensitive materials are prepared by dissolution in analytical grade chloroform. The solution is sprayed on the electrodes of the QCM from a tube with a very narrow tip (capillary). During coating, the frequency of the QCM is monitored and recorded, and after coating, each QCM sensor is compared with its standard sensor properties to avoid any errors from coating and synthesizing [11,12].

An approximation of the frequency change $f_k - f_{k+1}$ and gas mass change $m_k - m_{k+1}$ is given by the Sauerbrey equation (Eq. (1)) for consecutive samples k and $k+1$ [13]:

$$f_k - f_{k+1} = -\frac{C_f f_0^2}{A}(m_k - m_{k+1}) \tag{1}$$

where A is the area of sensitive layers, C_f is the mass sensitivity constant of the quartz crystal, f_0 is the fundamental resonance of the quartz crystals, and $(m_k - m_{k+1})$ are mass changes.

When a substance (such as a few hundred molecules of a target gas compound) is adsorbed onto these films, the resonant frequency of the quartz sensor changes. This frequency change is not permanent and can revert back to its original value when normal air is applied to the crystal surface (which is called cleaning or purging). Because of this reversibility, these devices are generally used as sensing elements for gases.

Sensor selection is an important task and requires expertise in both sensor development and sensor data processing [14,15]. An increase in the number of sensors in the cell does not necessarily guarantee an increase in information. For most systems, a number of sensors are developed for a target gas set, but only some of them are used in the e-nose because a few sensors may generate noisy data and some may have relatively insufficient sensitivity, or some sensors may be redundant due to similar dynamic behavior. In this research, 4 out of 9 sensors are selected, and their data are used for processing based on the factors mentioned above.

The e-nose used in this research is a device that is equipped with 9 different polymer film-coated quartz sensors and a reference sensor that is not coated with a steel cover on it. The e-nose generates numerical values that carry information about concentrations of the target gases at a second interval. The photo in Figure 1 shows the inside of the e-nose used for this research: there are 9 sensors in two rows, and a temperature and humidity sensor in the cell on the upper left corner (SHT11). The sensor cell is made of a hollow Teflon block.

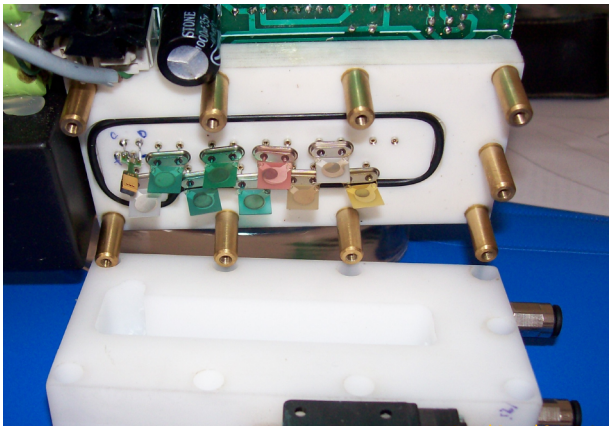

Figure 1. Inside the e-nose system: the sensor array and the sensor cell.

After the sensor array is cleaned (purged with dry air), a gas mixture is applied to the e-nose. The sensor frequencies change until they reach another stable state in which no more sensor frequency changes occur. This event takes from 5 to 15 min and is called a measurement. During a measurement, as many as 500 samples are collected from each sensor.

A steady-state response (Δf) is calculated subtracting the maximum value (at baseline) from the minimum value (at the balanced state) of sensor frequency responses as given in Eq. (2). The absolute value of the difference is taken since Δf is considered as the distance between the states (see Figure 2).

$$\Delta f = |f_{max} - f_{min}| \tag{2}$$

In this study, a population of steady-state data is obtained by conducting multiple measurements in series, which are then used for training the data processing system. Obtaining a data population is called an experiment and includes multiple measurements. Figure 3 shows the steady-state response points that are shaped like a surface

of a selected sensor during an experiment (Table 1). The x and y axes show incremental steps (500 ppm) for toluene and ethanol, and the z axis shows frequency shifts for each measurement. The surface is almost flat due to the linear sensor response in the range.

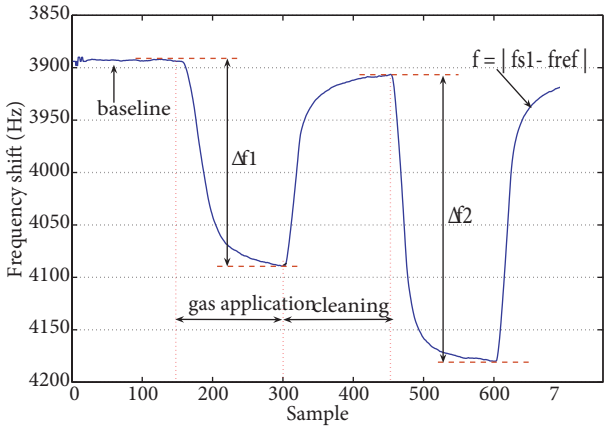

Figure 2. An example QCM sensor output: Δf_1 and Δf_2 are calculated using the dynamic samples as steady-state representatives for two different measurements.

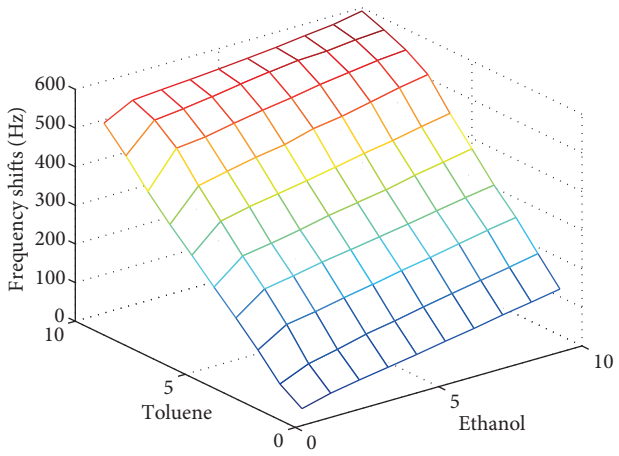

Figure 3. The steady-state responses of a selected sensor to variable concentrations of toluene and ethanol in Experiment C create a nearly flat surface.

3. The new sensor data evaluation method using PCA

Quantitative information extraction from gas sensor array data can be done using various linear and nonlinear methods. Since the relationship between inputs and outputs is multidimensional, a multiple regression method should be used to find a relation. Two of the well-known and commonly used multiple linear regression methods are listed as follows [16]: 1) the first order no-interaction model (FONI) of Eq. (3); and 2) the second order interaction model (SOI) of Eq. (4).

$$Y_{FONI-PCR} = b_0 + \sum_{i=1}^{k} b_i x_i \tag{3}$$

$$Y_{SOI-PCR} = (b_0 + \sum_{i=1}^{k} b_i x_i)(c_0 + \sum_{i=1}^{k} c_i x_i) \tag{4}$$

In these equations, x_i is the principal component scores (PC-i), Y_i is the corresponding gas concentration as ppm, b_i and c_i are regression coefficients, and k_i is the number of principal components (correspondingly used sensor count).

The sensor array data are transferred to the PCA domain to observe variances. After the operations over the training data, it is found that the PCs, except the first one, have smaller variances than 1% in total. Thus, PC-1 is found sufficient for concentration predictions. However, if the target is a mixture of gases, then PC-2 must be taken into account to determine the constituent species. The remaining correlated components could be disregarded. The input dimensionality of the new data processing system is reduced to 2 as an optimum solution. Since the components did not correlate in the PCA domain due to the orthogonalization procedure, the interaction terms drop out in Eq. (4). Then, the regression equations given above become as follows (see

Table 1. Training data for single and binary gas mixtures of toluene and ethanol. The first column indicates the experiment, the second column shows the measurement number, the third column shows gas concentrations, the fourth column lists principal components, and the last column shows the percent variances of the principal components in each experiment.

E. no.	M. no.	Concentrations (ppm)		PCs	Variances (%)
		Toluene	Ethanol		
A	1	0	0	PC-1	100.00
	2	0	512	PC-2	0.00
	3	0	1024	PC-3	0.00
	4	0	1536	PC-4	0.00
	5	0	2048		
	6	0	2560		
	7	0	3072		
	8	0	3584		
	9	0	4096		
	10	0	4608		
B	1	0	0	PC-1	99.66
	.	.	.	PC-2	0.33
	.	.	.	PC-3	0.00
	10	0	4608	PC-4	0.00
	11	490	0		
	12	1050	0		
	13	1540	0		
	14	2030	0		
	15	2520	0		
	16	3010	0		
	17	3500	0		
	18	4060	0		
	19	4550	0		
C	1	0	0	PC-1	99.65
	.	.	.	PC-2	0.31
	.	.	.	PC-3	0.04
	10	4550	0	PC-4	0.00
	11	490	512		
	.	.	.		
	.	.	.		
	19	490	4608		
	20	1050	512		
	.	.	.		
	.	.	.		
	37	1050	4608		
	.	.	.		
	.	.	.		
	.	.	.		
	100	4550	4608		

Eqs. (5) and (6)):

$$Y_{FONI-PCR} = b_0 + b_1 x_1 + b_2 x_2 \tag{5}$$

$$Y_{SONI-PCR} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 \tag{6}$$

Besides these well-known methods, we propose a new approach that simplifies calculations by reducing the independent variables to one without so much sacrificing from the prediction correctness. In this approach, a linear regression is used with only PC-1, and PC-2 is used along with PC-1 to determine constituent species (see Eq. (7)).

$$Y_{R-PCR} = b_0 + b_1 x_1 \tag{7}$$

The proposed approach is compared with regression methods given in Eqs. (5) and (6) in this work.

3.1. The proposed approach

The system comprises three major units: 1) a preprocessing unit; 2) a steady-state data extraction unit; and 3) a trainable PCR unit (see Figure 4). The preprocessing unit removes outliers caused by noise, and then filters quantization errors caused by the sampling system using a sliding window algorithm. Steady-state measurement data are calculated using dynamic sensor responses in the second unit.

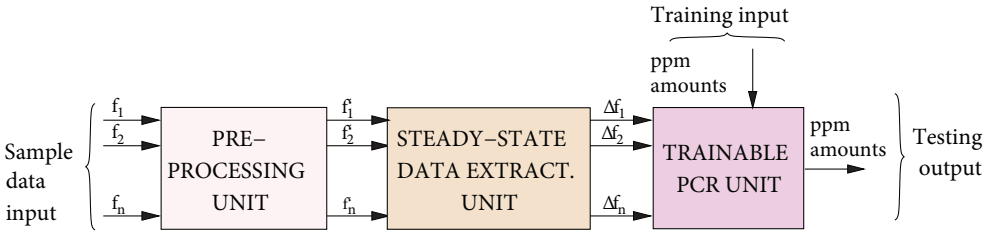

Figure 4. The data processing system is composed of three units: 1) the preprocessing unit removes outliers and filter quantization errors; 2) frequency shifts are calculated in the steady-state data extraction unit; and 3) arithmetic mean values, a transfer matrix, and approximation functions are determined in the training unit.

A line-fitting method is used for all training data to increase prediction quality. The predefined training sample space is shown in Figure 5. In the figure, PC-1 and PC-2 values of a sample point include the necessary

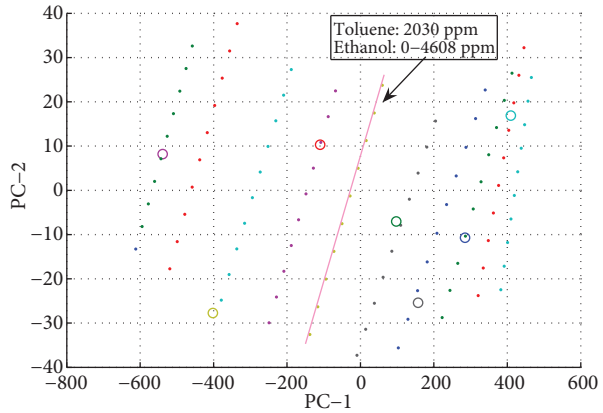

Figure 5. Training and testing samples used in the Experiment C. Each subpopulation forms a linear path, and test samples (shown with circles) may fall anywhere among the subpopulations.

information to predict the concentrations of the gases. When looking at the sample points in the PCA plot, it can be seen that the points create linear paths like lines. This is because we applied 1st order linear regression to the training samples.

The samples from a mixture with variable concentrations of a single constituent follow a linear path in the PCA plot. Consider an experiment conducted with measurements of a mixture that includes individual gases A and B, such that either the concentration of A is fixed and that of B is changed, or the concentration of B is fixed and that of A is changed. The PCA plot of this experiment includes two linear paths: the fixed concentration amounts are represented by lines, and the variable amounts are represented by points on the lines (see Figure 5). If an experiment includes all possible balanced steps, then the sample space becomes a surface knit by lines.

The trainable PCR unit forms a discrete sample space with the supplied data. As more data are provided during training, prediction accuracy increases in testing. The unit contains memory to save training parameters for each experiment. These parameters are: 1) arithmetic mean values; 2) a transfer matrix; and 3) a matrix for line functions.

3.2. Calculating training parameters

The training parameters are obtained using the steady-state sensor responses (Δf), which are calculated using collected dynamic samples from the selected sensors. The sensor sample populations are formed conducting several measurements during an experiment.

Multiple measurements of each sensor in an experiment are saved into memory as vectors of Δfs. This vector is known as a sample vector $(\mathbf{SV}_{m\times 1})$ as in Eq. (8). For example, $\mathbf{SV}_1$ and $\mathbf{SV}_2$ represent measurement vectors for sensor-1 and sensor-2, respectively.

$$\mathbf{SV}_i = \begin{bmatrix} \Delta f_1 & \Delta f_2 & \dots & \Delta f_m \end{bmatrix}^T \; i = 1, \dots, n \tag{8}$$

where n and m represent sensor count and measurement count, respectively. In this work, $n = 4$, and $m = 10$, which is experiment-dependent.

A matrix, called the sample matrix $(\mathbf{SM}_{m\times n})$, is formed using the $\mathbf{SV}$s for each sensor in the system. The rows of this matrix represent measurements, and the columns represent sensor responses.

$$\mathbf{SM} = \begin{bmatrix} \mathbf{SV}_1 & \mathbf{SV}_2 & \dots \mathbf{SV}_n \end{bmatrix} \tag{9}$$

The following list details how to obtain the training parameters:

1. **The arithmetic mean values:** Arithmetic averages are calculated for each column of $\mathbf{SM}$ in order to center the data. Arithmetic mean values can be arranged as a vector, as shown in Eq. (10).

$$\mu_i = E\left(\mathbf{SV}_i\right) \tag{10}$$

$$\mu\mathbf{V} = \begin{bmatrix} \mu_1 & \mu_2 & \dots \mu_n \end{bmatrix}^T \tag{11}$$

2. **The transfer matrix:**

   - A covariance matrix $(\mathbf{C}_{n\times n})$ is calculated using the $\mathbf{SM}$:

$$\mathbf{C} = cov\,(\mathbf{SM}) \tag{12}$$

- Eigenvalues and eigenvectors of this covariance matrix are calculated using the following equations:

$$det\,[\lambda\mathbf{I} - \mathbf{C}] = 0, \lambda_i sareeigenvalues, i = 1, \ldots, n. \tag{13}$$

$$\mathbf{C}\mathbf{X}_i = \lambda_i \mathbf{X}_i, \mathbf{X}_i istheeigenvectorof\lambda_i. \tag{14}$$

- A transfer matrix ($\mathbf{TM}_{n\times n}$) is formed using eigenvectors ($\mathbf{X}_i$s) in columns, with the most significant eigenvector in the first column.

$$\mathbf{TM} = \left[\begin{array}{cccc} \mathbf{X}_1 & \mathbf{X}_2 & \ldots & \mathbf{X}_n \end{array}\right]^T \tag{15}$$

The transfer matrix is used both in training and testing. In training, all data in the population are transferred to the PC domain at once using the following equation:

$$\mathbf{ASM}_{n\times m} = \left[\begin{array}{ccc} \mathbf{SV}_1 - \mu_1 & \cdots & \mathbf{SV}_n - \mu_n \end{array}\right]^T \tag{16}$$

$$\mathbf{TSM}_{n\times m} = \mathbf{TM}_{n\times n} \times \mathbf{ASM}_{n\times m} \tag{17}$$

where **TSM** denotes a transferred sample matrix (PCA scores).

3. **The line functions:** The transferred samples in the training phase form a new domain: the principal component domain. In this domain, when the concentration of one constituent component of a mixture is held fixed and the other is changed incrementally, special subpopulations occur in the total population. Linear paths (lines) known as training lines (see Figure 6) appear in the domain and their number corresponds to the number of special subpopulations (l). A first-order polynomial regression function is obtained using training samples on each linear path. All training line functions produce a training function vector(f_i), as shown in Eq. (18).

$$\mathbf{FV}_{training} = \left[\begin{array}{ccc} f_1 & f_2 & \cdots f_l \end{array}\right]^T \tag{18}$$

A linear relationship is observed between PC values and the concentrations during preliminary experiments. Since PC-1 is the most informative component of the PCA, we decided to select only PC-1 for the regression. The first order polynomial regression that describes this relation is called the transition function, and the line that represents this function is called the transition line, as shown in Figure 7. The transition line function vector (g_i) is formed by all transition line functions given in Eq. (19).

$$\mathbf{FV}_{transition} = \left[\begin{array}{ccc} g_1 & g_2 & \cdots g_l \end{array}\right]^T \tag{19}$$

Training line and transition line functions are used to create a matrix where the first column holds the training data and the second column contains transition line functions, as in Eq. (20).

$$\mathbf{FM}_{l\times 2} = \left[\begin{array}{cc} \mathbf{FV}_{training} & \mathbf{FV}_{transition} \end{array}\right]^T \tag{20}$$

Training line functions compose the training space. The row index of **FM** points to the line functions, so traversing in the training space is fairly straightforward. These functions are formed in the training phase, and they are used in the testing phase.

3.3. Using the system for a test sample

In a measurement, a steady-state sample vector (**TSV**) that comes from sensor array is converted to an adjusted test sample vector (**ATSV**) by subtracting the arithmetic mean vector (μ**V**), a training parameter for the system (see Eq. (21)). Then, this modified test sample is transferred to the PC domain using Eq. (22). A transferred test sample is shown in Figure 6.

$$\mathbf{ATSV} = \mathbf{SV} - \mu\mathbf{V} \tag{21}$$

$$\mathbf{TTSV} = \mathbf{TM} - \mathbf{ATSV} \tag{22}$$

A test sample point may fall in a location where no training line crosses it, and cannot be recognized in the training unit. Thus, a virtual line intersecting the test sample point can be formed using regression analysis for the prediction (see Figure 7). However, this requires an additional virtual line (transition) to convert a PC-1 value to an amount in ppm. Another solution is to merge the sample point into the closest subpopulation (training line). Since it is simpler, requires less calculation, and results in reasonable error, the second solution is preferred. The tasks to find quantitative outputs for a measurement sample can be listed as follows:

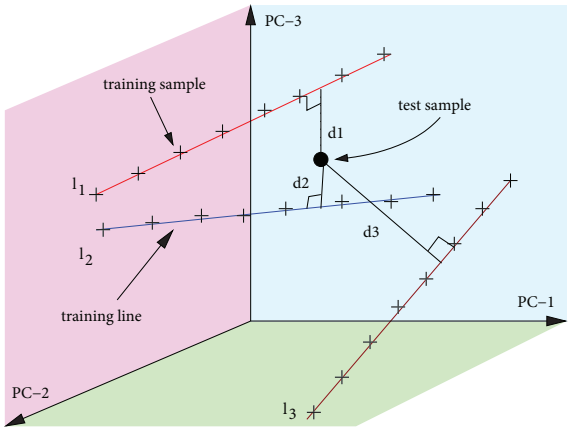

Figure 6. Finding the closest line to the test sample point in a 3D PC domain.

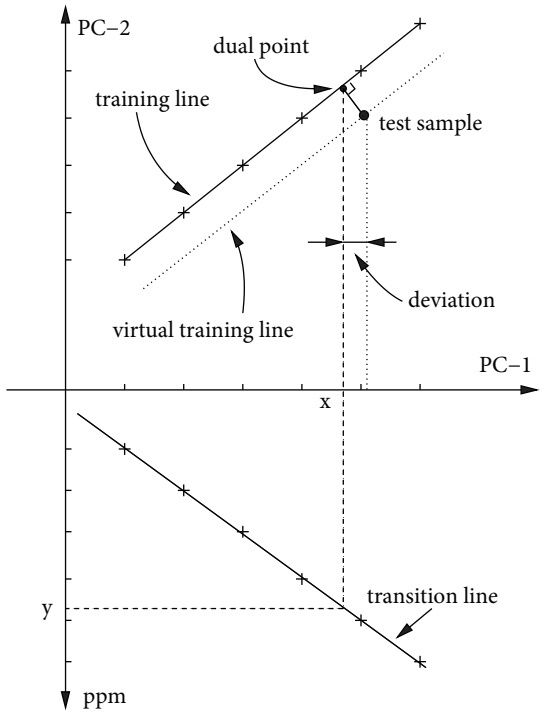

Figure 7. The graph shows how a concentration amount is obtained from a test sample in the PC domain.

1. Find the closest line to the test sample: when the test sample is associated with a line in the space, a numerical value related to the concentration amount of the first constituent component can be found for

that test sample. Searching a line that has a minimum distance to the sample point is conducted by calculating the perpendicular distances to each line using Eq. (23). Figure 6 shows how to find the closest line to a sample point graphically.

$$d_i = \frac{|(\mathbf{X}_{i,2} - X_{i,1}) \times (\mathbf{X}_{i,1} - \mathbf{sp})|}{|\mathbf{X}_{i,2} - \mathbf{X}_{i,l}|}, i = 1, \ldots, l. \tag{23}$$

where $\mathbf{X}_{i,j}$ represents the jth point on the ith line, **sp** represents the sample point, and l is the line count.

2. Find the closest point (dual point) on the line: the dual point is the footprint of the test sample on the closest training line. The conversion process from PC values to concentration (ppm) is shown graphically in Figure 7. The transition line on the figure is obtained easily following the steps of the approach. The maximum absolute error in this method equals one half of a step (approximately 250 ppm for this study). Therefore, using sample data, the concentration is predicted conveniently by doing some linear matrix operations. Figure 7 illustrates concentration prediction in a two-dimensional PC domain.

MATLAB software by MathWorks was used for all calculations.

4. Experimental setup

Figure 8 shows the experimental setup used to obtain samples for the training and testing stages. Dry air in a tank was used to clean sensors before gas application to the chamber. A microcontroller system (MKS unit) that generates programmed electrical signals was used to control the valves (MFC: mass flow controller) to obtain desired flow rate. Software run on a computer was developed to implement time-based tasks in the MKS. The other software run on the same computer receives dynamic samples every second and saves them to files on the internal disk. Both the MKS unit and the e-nose communicate with the computer over the serial ports (RS-232).

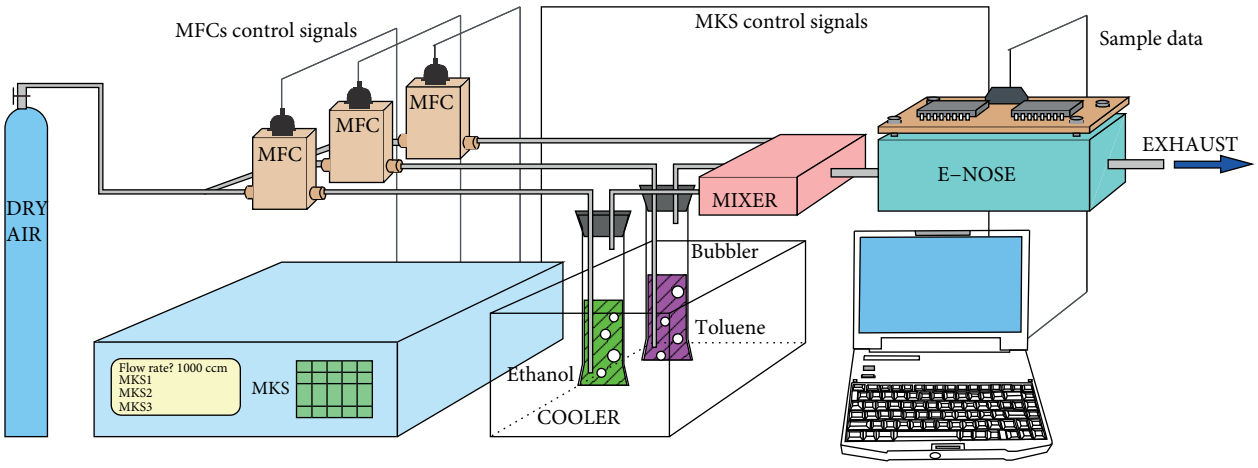

Figure 8. The experimental setup to obtain the desired gas concentrations from toluene and ethanol.

Condition parameters such as temperature and flow rate are calculated using Antoine's equation to obtain the desired gas concentrations [15]. Individual gases are obtained by adding toluene and ethanol to dry air using

a bubbler in a temperature-controlled cabin. Then, the individual gases are mixed together in another chamber (the mixer).

During the experiments, single and binary gas mixtures are obtained. For this work, the MKS unit is programmed to obtain a 0–4500 ppm range of concentrations for constituent components with approximately 500 ppm steps. Table 1 shows some of the experimental data presented in this paper.

Three different experiments were selected from a series of extensive tests with various scenarios for two industrial gases: toluene and ethanol. The test experiments were designed as follows:

- Experiment A utilized a single type of gas (ethanol) in variable concentration measurements. The performance evaluation of a single gas concentration prediction is the goal of this experiment.

- Experiment B makes measurements on two different gases with nonoverlapping variable concentrations. The total population is formed from two subpopulations, one of which comes from Experiment A.

- Experiment C is a comprehensive test: it covers all of the prior experiments, and forms a composite population. The transferred space is expected to answer for all possible input test vectors of single and binary mixtures of the target gases.

5. Results and discussion

In this section, test results concerning the system performance are reported. The system was initially trained separately using three experiments, as shown in Table 1. Then, the trained system was tested with random measurements, as shown in Table 2.

Table 2. Test results of the approaches (R-PCR, FONI-PCR, and SONI-PCR) for each experiment in Table 1.

		Real values		Predictions (ppm)						Error (%)					
E.	M.	Conc. (ppm)		R-PCR*		FONI-PCR		SONI-PCR		R-PCR*		FONI-PCR		SONI-PCR	
no.	no.	Tol.	Eth.	Tol.	Eth.	Tol.	Eth.	Tol.	Eth.	Tol.	Eth.	Tol.	Eth.	Tol.	Eth.
A	1	-	2200	-	2216	-	2216	-	2216	-	0.73	-	0.73	-	0.73
B	1	-	3700	-	3713	-	3521	-	3777	-	0.35	-	4.84	-	2.08
	2	1100	-	1095	-	1109	-	995	-	0.45	-	0.82	-	9.55	-
C	1	-	2200	-	2193	-	2517	-	1888	-	0.32	-	14.4	-	14.2
	2	1100	-	1050	-	855	-	950	-	4.54	-	22.3	-	13.6	-
	3	2900	910	3010	929	3217	505	3181	582	3.79	2.09	10.9	44.6	9.68	36.1
	4	3400	1420	3500	1434	3607	1710	3630	1595	2.94	0.99	6.08	20.4	6.75	12.3
	5	2440	2600	2520	2814	2767	1846	2587	2217	3.28	8.23	13.4	29.0	6.01	14.7
	6	1540	3584	1540	3531	1697	3000	1440	3559	0.00	1.48	10.2	16.3	6.47	0.69
	7	4060	3072	4060	2818	3844	3887	3954	3677	0.00	8.27	5.32	26.5	2.60	19.7
* Reduced-PCR (our approach).								Mean error		2.14	3.10	9.86	22.3	7.81	14.3
								Max. error		4.54	8.27	22.3	44.6	13.6	36.1

Experiment A: *The system was trained with 10 samples of ethanol with variable concentrations, and tested with a random sample. PC-1 carries almost all informational data, so the training line lies on the PC-1 axis. Therefore, only the PC-1 value of a test sample and the transition line were used for predictions. Images of the training and test samples over the transition line for Experiment A are shown in Figure 9. The system predicted the concentration amount with a very high success rate (99.29%).*

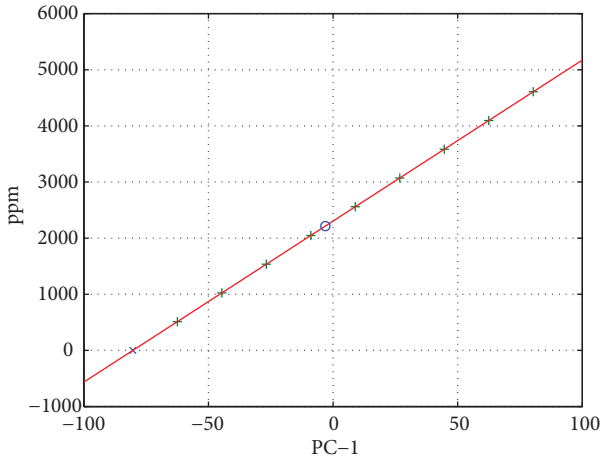

Figure 9. PC-1 versus concentration amounts for Experiment A.

Concentration amounts for this experiment could also be calculated from steady-state responses without transferring samples to the PC domain. However, knowing which sensor(s) responses should be used for a prediction is difficult. PCA provides a linear solution to this problem by reducing the dimension of all sensor responses (i.e. using PC-1 alone is sufficient to find the ppm amounts).

Experiment B: *The system was trained with 19 samples of toluene and ethanol, and tested with 2 random samples of each species. There were two subpopulations (that form two different linear paths) in this experiment, as shown in Figure 10a. First, it must be determined which linear path is closest to the test sample, and then the steps of the approach must be followed to find out the concentration amounts, as was explained in detail in Section 3.3.*

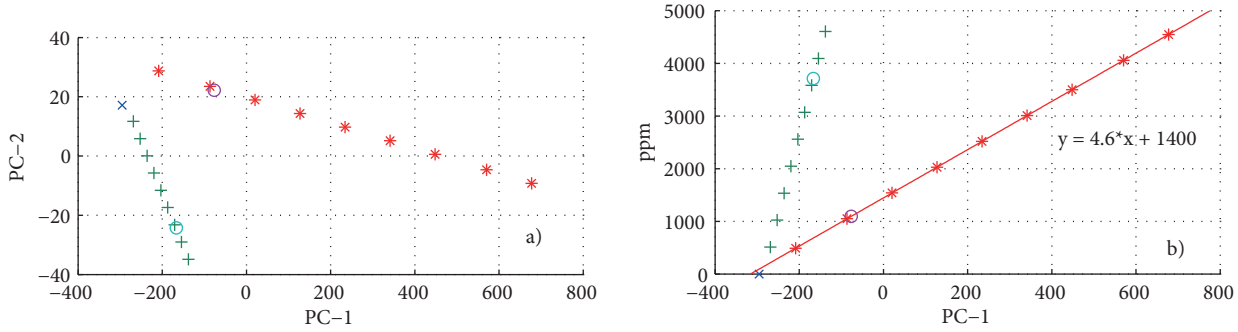

Figure 10. a) PC-1 versus PC-2; and b) PC-1 values in a) versus concentrations. The plus and star symbols represent two subpopulations (ethanol and toluene respectively), and circles are used for the test samples.

Figure 3 shows that increasing gas concentrations are related linearly with steady-state sensor responses (Δfs). More than 99% of the information that these responses contain relies on the first principal component of the PC domain. Figure 10b shows the transition lines used to convert PC-1 values to concentrations in ppm in this experiment. The transition lines shown in Figure 10b are associated with the training paths given in Figure 10a.

At least two principal components must be used to predict concentration amounts in this experiment since there are two different species. As presented in Table 2, the percent error rate of the system was less than 1% for both test samples.

Experiment C: *The system was trained with 100 samples and tested with seven samples. Training data for this experiment create many discrete lines in the PC domain (see Figure 4). The test samples are selected as follows: 1) the first group (two samples) is the same as in Experiment A and B, such that these points lies on different lines, but do not lie on any training points; 2) the second group (three samples) is selected randomly such that they all fall along lines at different locations; and 3) the last group (two samples) is selected from the training samples in experiment C given in Table 1.*

The first principal component holds 99.65% of the concentration data. The PC-1 and PC-2 components together create a training surface, such that the points on each training line contain the ethanol concentration data, and the location of each training line contains the concentration data for toluene in the mixture. Numerical values for the training line are shown in Figure 4 as an example. Two principal components are found to be sufficient to analyze a mixture that contains two species and their binary mixtures.

Concentration amounts for the test samples were predicted using three methods: R-PCR, FONI-PCR, and SONI-PCR. The results are presented in Table 2. The R-PCR method performed better than the other two for all tests, since PC-2 contained very low variance compared with PC-1. Hence, detailed explanations about the R-PCR method are listed as follows:

- Experiment A-1 and Experiment C-1 were done with the same test sample of ethanol (2200 ppm). The results are reasonable with a very low rate of error.
- Experiment B-2 and Experiment C-2 were done with the same test sample of toluene (1100 ppm). The error rate of measurement in C-2 (4.54%) is greater than that in B-2 (0.45%). This occurs as a result of merging the test sample to the nearest subpopulation (line).
- The rates of error for measurements C-2, C-3, and C-4 are within a reasonable range.
- The rates of error for toluene in measurements C-6 and C-7 are 0.00% because the points were selected from the training samples. Thus, merging the test point to the closest line results in zero error.

However, ethanol predictions do not show similar behavior. In measurement C-6, the predicted ethanol concentration deviated from the real value. This occurred because the line fit was applied to the steady-state responses in the preprocessing stage. The training points on the line deviated from the original locations, resulting in an error measurement of 1.48%.

For the same reason, ethanol prediction also deviated in measurement C-7. The deviation amount is greater than that in the previous measurement. The previous sample fell in the almost pure linear region of the steady-state response surface, as shown in Figure 2. However, the samples were selected from a spot where the linearity of the surface was relatively low. Hence, the error rate was found to be 8.27%.

The outcomes of the test samples are predicted by approximating them to the closest line, and to the closest point on the line. Additional training data create a better space with more lines, resulting in a lower absolute prediction error.

Finally, test prediction deviations were calculated and put into Table 2. The maximum possible absolute error is equal to half of a step for the R-PCR method. The maximum possible relative error in this case is 50% for measurements near the lower concentrations (i.e. 500 ppm), and 5.6% near the last step (i.e. 4500 ppm) in the R-PCR method.

The amount of possible maximum absolute error could be reduced by introducing more training data, but again, the possible maximum relative error values above would stay the same. For example, instead of using

steps of 500 ppm, steps of 100 ppm could be used, and the maximum absolute error would drop from 250 ppm to 50 ppm. Moreover, instead of merging a point to a linear path, a virtual linear path could be formed using a surface fit approach in the domain. This would lower the error rates for the toluene predictions (compare the toluene percent errors in Experiment B-2 and in Experiment C-2).

Although the test points were selected from different locations in the space, the maximum error rate was calculated as 8.27% in this study for the R-PCR method. These low rates of error are promising, and can be lowered still by providing more data in the training stage.

- This approach presents a flexible gas measurement system that can be used for many different target gas sets without changing the hardware. The system must be trained for a target gas set before it is ready for use. No hardware changes are required if the sensor array holds a sufficient number of sensor types.
- Compared with nonlinear approaches (such as artificial neural networks), this approach makes it easier to implement handy instruments for gas measurement, since only linear operators are used for calculations (addition and multiplication, with no exponential operator required).
- Single and binary mixtures of two gases are used in this work. The study can be extended to include more species in the mixture. In that case, the number of sensors in the array and the number of principal components must be increased.

6. Conclusion

This paper presented a novel use of PCR for finding the concentration amounts of the constituent components of a mixture. This approach provides three desired properties of quantitative gas evaluation systems: retrainability, flexibility, and simplicity. With these properties, a system can be trained for different target gas sets and then used for measurements without any hardware modification. The method is a linear approach; hence, it does not require powerful computation resources as do nonlinear approaches. The system was tested with individual and binary mixtures of toluene and ethanol, and promising results were obtained. This approach can easily be implemented in micro-PC or microcontroller based embedded systems for portable gas measurement devices.

References

[1] Özmen A, Tekçe F, Ebeoğlu MA, Tasaltın C, Öztürk ZZ. Finding the composition of gas mixtures by a phthalocyanine coated QCM sensor array and an artificial neural network. Sensor Actuat B-Chem 2006; 115: 450–454.

[2] Mumyakmaz B, Özmen A, Ebeoğlu MA, Taşaltın C. Predicting gas concentrations of ternary gas mixtures for a predefined 3-d sample space. Sensor Actuat B-Chem 2008; 128: 594–602.

[3] Kermit M, Tomic O. Independent component analysis applied on gas sensor array measurement data. IEEE Sens J 2003; 3: 218–228.

[4] Polikar R, Shinar R, Udpa L, Porter M. Artificial intelligence methods for selection of volatile organic compounds. Sensor Actuat B-Chem 2001; 80: 243–254.

[5] Jolliffe IT. Principal Component Analysis. 2nd ed. New York, NY, USA: Springer, 2002.

[6] Aleixandre M, Lozano J, Gutierrez J, Sayago I, Fernandez MJ, Horrillo MC. Portable e-nose to classify different kinds of wine. Sensor Actuat B-Chem 2008; 131: 71–76.

[7] Dutta R, Hines EL, Gardner JW, Kashwan KR, Bhuyan M. Tea quality prediction using a tin oxide-based electronic nose: an artificial intelligence approach. Sensor Actuat B-Chem 2003; 94: 228–237.

[8] Jolliffe IT. A note on the use of principal components in regression. J Roy Stat Soc C-App 1982; 31: 300–303.

[9] Mumyakmaz B., Yamaçlı M. A nonlinear principal component regression application: long-term electrical load demand forecasting of Turkey. In: 1st International Symposium on Computing in Science & Engineering; June 2010; Aydın, Turkey. pp. 537–541.

[10] Mumyakmaz B, Özmen A, Ebeoğlu MA, Taşaltın C, Gürol İ. A study on the development of a compensation method for humidity effect in QCM sensor responses. Sensor Actuat B-Chem 2010; 277–282.

[11] Gürol İ, Ahsen V, Bekaroğlu Ö. Synthesis of tetraalkythio-substituted phthalocyanines and their complexation with Ag^{I} and Pd^{II}. J Chem Soc Dalton Trans 1994; 497–500.

[12] Gürol İ, Ahsen V. Synthesis and complexation of a new soluble multidentate diaminoglyoxime derivative. Syn React Inorg Met 2001; 31: 127–138.

[13] King HW. Piezoelectric sorption detector. Anal Chem 1964; 36: 1735–1739.

[14] Boilot P, Hines EL, Gongora MA, Folland RS. Electronic noses inter-comparison, data fusion and sensor selection in discrimination of standard fruit selections. Sensor Actuat B-Chem 2003; 88: 80–88.

[15] Gardner JW, Boilot P, Hines EL. Enhancing electronic nose performance by sensor selection using a new integer-based genetic algorithm approach. Sensor Actuat B-Chem 2005; 106: 114–121.

[16] Schiff D, D'agostino RB. Practical Engineering Statistics. New York, NY, USA: Wiley, 1996.

[17] Riddick JA, Bunger WB. Techniques of Chemistry. New York, NY, USA: Wiley, 1970.